Genomic Identifier CDEs as candidate data standards

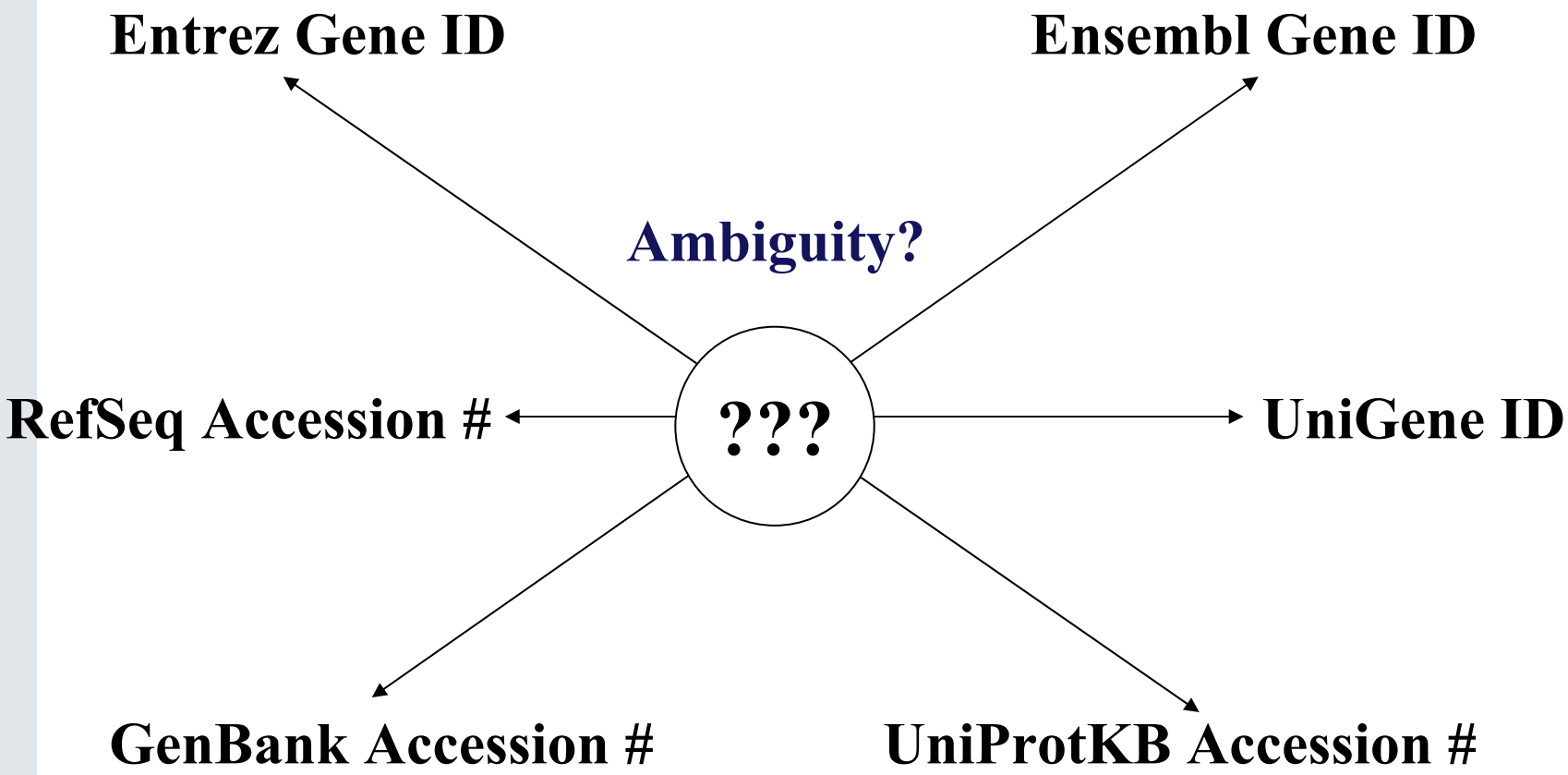
Craig Street/ Vishal Nayak
Abramson Cancer Center
University of Pennsylvania

The Problem?

Lack of a **COMMON** genomic identifier in biomedical databases

Results in an **INTEROPERABILITY** problem

The Problem?

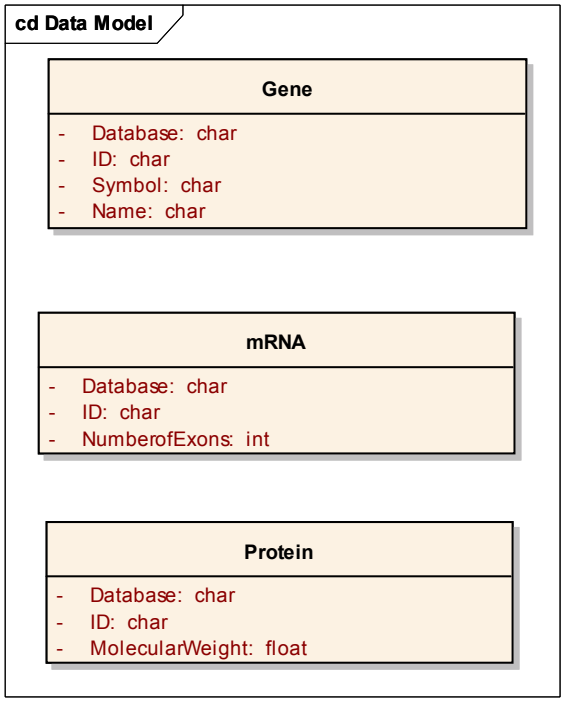

Entrez Gene ID
Ensembl Gene ID
Ambiguity?
RefSeq Accession #
???
UniGene ID
GenBank Accession #
UniProtKB Accession #

Initial Proposal

- Identify potential genomic identifiers from current ICR projects and potential future use cases
- Reuse existing CDEs in the caDSR and create new CDEs for these genomic identifiers
- Each ICR project's data model should utilize at least one or more of these defined CDEs
- Additional CDEs may be added to this dynamic list
- The Architecture Workspace, Vocabulary/CDE Workspace, and Genome Annotation subject matter experts should oversee the addition of such CDEs

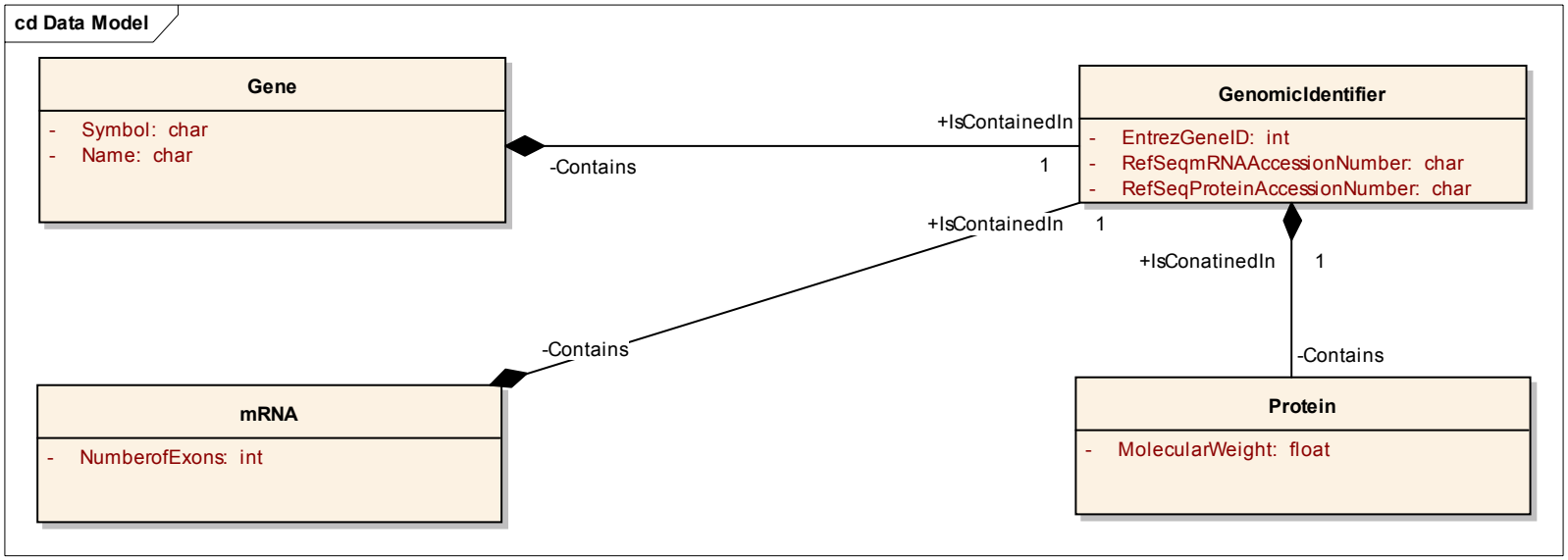
Evolution of the Proposal

- Proposal: Gene, RNA, and Protein classes each have an identifier and a provided database identifier rather than as a genomic identifier itself



Evolution of the Proposal

- **Proposal: Object-Relational mapping will be applied**



Genomic Identifier Modeling Strategy

- Polled Developers of ICR Projects to Determine What Identifiers are Currently Being Used in their Systems
- Compiled List of Required Genomic Identifiers (i.e., DNA or its RNA or Protein Product) based on Survey
- Each ICR Project's Object Model must utilize AT LEAST ONE of these Defined Genomic Identifier CDEs for each UML Class with any Genomic Identifier Attributes.

Genomic Identifier Modeling Strategy

- **If Object Models in the Future Cannot Accommodate one of the Existing CDEs, Additional CDEs may be Added to this Dynamic List.**
- **To Increase Unification in the Grid, in addition to their Use in UML Models, we Recommend that Applications Accept and Return as Parameters the Defined Genomic Identifier CDEs in their APIs.**

Genomic Identifier Modeling Strategy

List of Recommended Identifiers

- Ensembl Gene ID
- Ensembl Transcript ID
- Ensembl Peptide ID
- Gene GenBank Accession Number
- mRNA GenBank Accession Number
- Gene GenBank Accession.Version
- mRNA GenBank Accession.Version
- Entrez Gene ID
- RefSeq mRNA Accession Number
- RefSeq Protein Accession Number
- UniGene Cluster ID
- UniProtKB Primary Accession Number

Genomic Identifier Modeling Strategy

Genomic Identifier Property:

- **Define Separate Identifier Term for Genomics Classes**
- ***Genomic Identifier* Imparts more Specificity and Distinguishes from other Type of Identifiers (e.g., patients, tissues, etc.)**
- **Will be Advantageous when Querying across the Grid - Defining what Type of Data to Return**
- **Provides a Means of Interlinking Various Object Models**

Genomic Identifier Modeling Strategy

An Illustrative Example

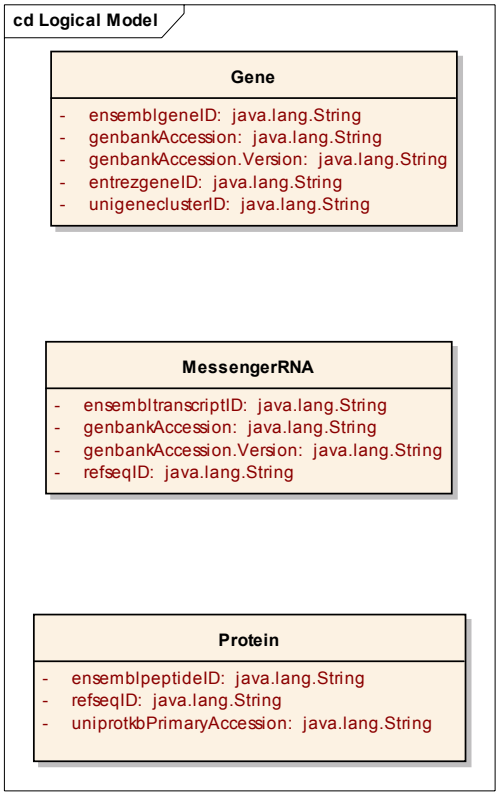
The following example shows how one would query disparate object types mapped with the same identifier:

Entrez Gene ID 672 BRCA1: breast cancer 1, early onset

Washington University
CDE-->DEC--->Object Class = Ribonucleic Acid (EVS concept code: C812)
CDE-->DEC--->Class Qualifier = Messenger RNA (EVS concept code: C813)
CDE-->DEC--->Property = Genomic Identifier (EVS concept code: C45766)
CDE-->DEC--->Property Qualifier = Entrez Gene (EVS concept code: C45765)
CDE-->VD--->Syntax constrained to Entrez Gene identifiers, but not enumerated

Georgetown University
CDE-->DEC--->Object Class = Protein (EVS concept code: C17021)
CDE-->DEC--->Class Qualifier = *optional (one could specify the protein family)*
CDE-->DEC--->Property = Genomic Identifier (EVS concept code: C45766)
CDE-->DEC--->Property Qualifier = Entrez Gene ID (EVS concept code: C45765)
CDE-->VD--->Syntax constrained to Entrez Gene identifiers, but not enumerated

UML Model - Class Diagram



Genomic Identifier Modeling Strategy

Alternate Modeling Approach

- **Specifically Designed for Projects that Cannot Accommodate the Recommended Strategy**
- **Essentially, this Approach Requires Objects which are "Served Out" and Reference Genes or Gene Products Must Contain an Embedded Identifier.**

Genomic Identifier CDE Creation

Current Status

- **The Genomic Identifier CDEs have been loaded into the caDSR production server.**
- **The Data Elements (including the Value Domains) have undergone one round of curation.** ***(courtesy Tommie Curtis, Claire Wolfe, Hong Dang, Craig Street and yours truly.)***
- **All definitions used have been validated using either an authoritative source or citation.**

Genomic Identifier CDE Creation

Current Status

- **Reference Documents have been attached to the Genomic Identifier CDEs.**
- **The Genomic Identifier CDEs have undergone an initial review by members of Genome Annotation SIG.**
- **The CDE creation process has been documented and checked into CVS.**
  (http://cabigcvs.nci.nih.gov/viewcvs/viewcvs.cgi/pir/adopter_penn/month3/Documentation%20of%20Genomic%20Identifier%20CDEs.doc)

Genomic Identifier CDE Creation

Next Steps

- **Get the curated Genomic Identifier CDEs reviewed by the VCDE and the members of the Genome Annotation SIG.**
- **Incorporate gathered feedback.**
- **Propose the Genomic Identifier CDEs as Candidate Data Standards.**

Genomic Identifier CDEs

- Ensembl Gene ID
- Ensembl Transcript ID
- Ensembl Peptide ID
- Gene GenBank Accession Number
- mRNA GenBank Accession Number
- Gene GenBank Accession.Version
- mRNA GenBank Accession.Version
- Entrez Gene ID
- RefSeq mRNA Accession Number
- RefSeq Protein Accession Number
- UniGene Cluster ID
- UniProtKB Primary Accession Number

Why Data Standards?

- These data elements are very commonly used within the ICR workspace.
- Promote reuse of the CDEs to ease mapping of the genomic identifiers.
- Prevent post harmonization issues by encouraging developers to harmonize their data elements on these CDEs during the development process.
- Encourage the development of more commonly used CDEs within the ICR workspace as Data Standards.
- Washington University has been funded in Year 2 to create an identifier mapping service based on this list of identifiers, to further support this approach.